Data Infrastructure for Massive Scientific Visualization and Analysis

James Ahrens & Christopher Mitchell

Los Alamos National Laboratory

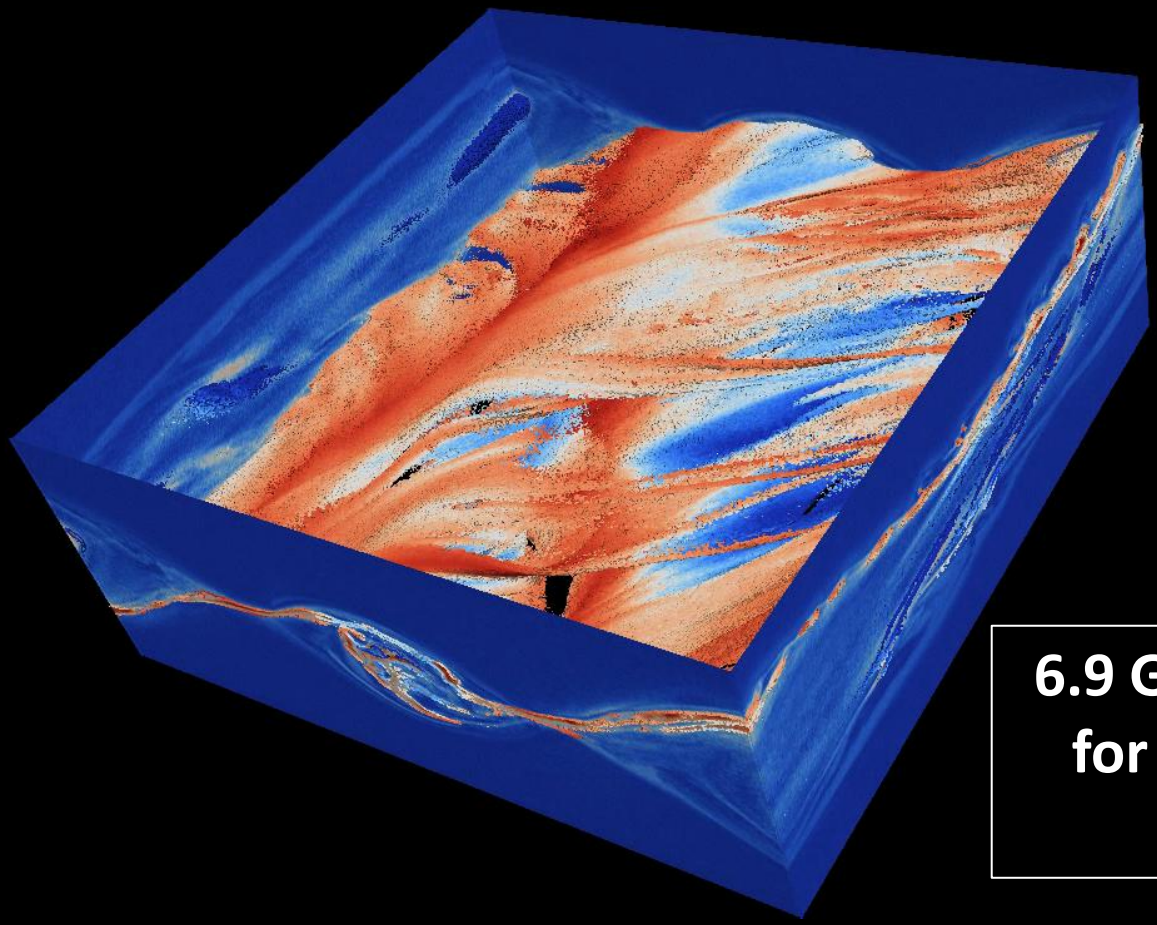
Large Data in High Performance Computing

- HPC users are simulating the world to better understand it and the products we use.
  - Climate, High Energy Physics, Defense Problems, etc.
- Current simulations are detailed enough to produce hundreds of TB if not PB of data.
- Raw data is useless without analysis and/or visualization.

A Scalable Open-source Visualization and Analysis Tool - ParaView

- ParaView – *http://www.paraview.org*
  - Integrated analysis and visualization platform for high-performance computing
  - Data Parallelism
    - Works well due to massive size of each dataset
  - Data Streaming
    - Incrementally process what is requested
  - Multi-Resolution
    - Immediate reduced resolution results
      - Full resolution streams in over time

Example: Plasma Data in ParaView



6.9 GB/time step for one vector field

Integration with Data Intensive Technologies

- Project initiated to rethink I/O in ParaView and storage for HPC analytics.
- Currently use Parallel File System to access data via a storage network
- Data too expensive to ship from storage to processor
  - No longer “interactive” to user
- Rewrote ParaView reader to interface with the Hadoop Distributed File System.
- Process data *in situ* to where it is stored.
- Key is scheduling algorithm
  - Match nodes with data to ParaView processes
  - *No network transfers!*

Integration with Data Intensive Technologies

- Seeing near *3x improvement* in read times compared to existing solutions.
  - What previously took ~1.5 seconds to load can be done in less than 0.5 seconds per time step!
- Lower standard deviation on results
  - More consistent read times.